BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC KINH TẾ QUỐC

---------\*\*\*--------



DỰ BÁO CHUỖI DỮ LIỆU PHỤ THUỘC THỜI GIAN BẰNG MÔ HÌNH ARIMA

Group 10

| No1 | ID | Full name | Contribution for common section |
| --- | --- | --- | --- |
| 1 | 11208006 | Đỗ Thùy Trang |  |
| 2 | 11202127 | Hoàng Diệu Linh |  |
| 3 | 11201426 | Nguyễn Thị Thanh Hiền |  |
| Sum |  |  | 100% |

Hà Nội, 2023

**INTRODUCTION**

To predict an outcome based on time series data, we can use a time series model which is called Auto Regressive Integrated Moving Average (ARIMA). It is used as the machine learning technique to analyze and predict future stock prices based on historical prices.

ARIMA (autoregressive integrated moving average) is a commonly used technique utilized to fit time series data and forecasting. It is a generalized version of ARMA (autoregressive moving average) process, where the ARMA process is applied for a different version of the data rather than the original. Three numbers p, d and q specify ARIMA model and the ARIMA model is said to be of order (p, d, q). Here p, d and q are the orders of AR part, Difference, and the MA part respectively. AR and MA- both are different techniques for stationary time series data. ARMA (and ARIMA) is a combination of these two methods for better fitting the model. In this write-up an overview of the AR and MA process will be given. The steps of building an ARIMA model will be explained. Finally, a demonstration using R will be presented.

1. Individual
2. Đỗ Thùy Trang
3. CTCP Vinacafé Biên Hòa (VCF: HOSE)

Vinacafé BH inherited the biggest achievement of Bien Hoa Coffee Factory, formerly the Vinacafé brand. Established in the 1980s and officially recognized as an intellectual property in 1993, Vinacafé brand today has become a major brand of Vietnam, selected into the National Brand Program since 2008. Vinacafé painstakingly built from a solid foundation: product quality and commitment "Taste of nature."

1. Finance series

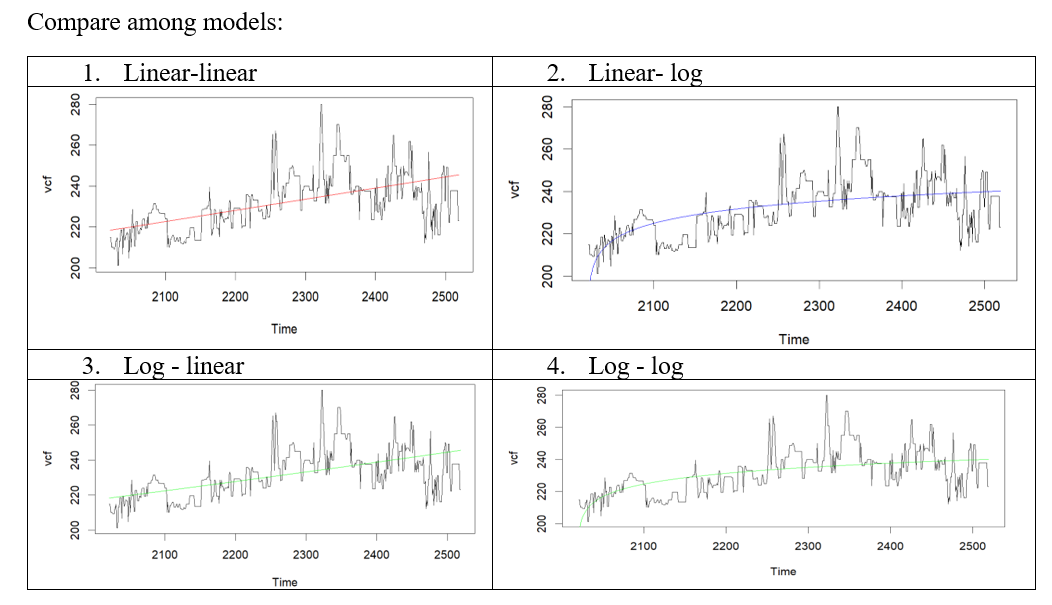
Daily\_data : <close\_price> of VCF in 2021-2022

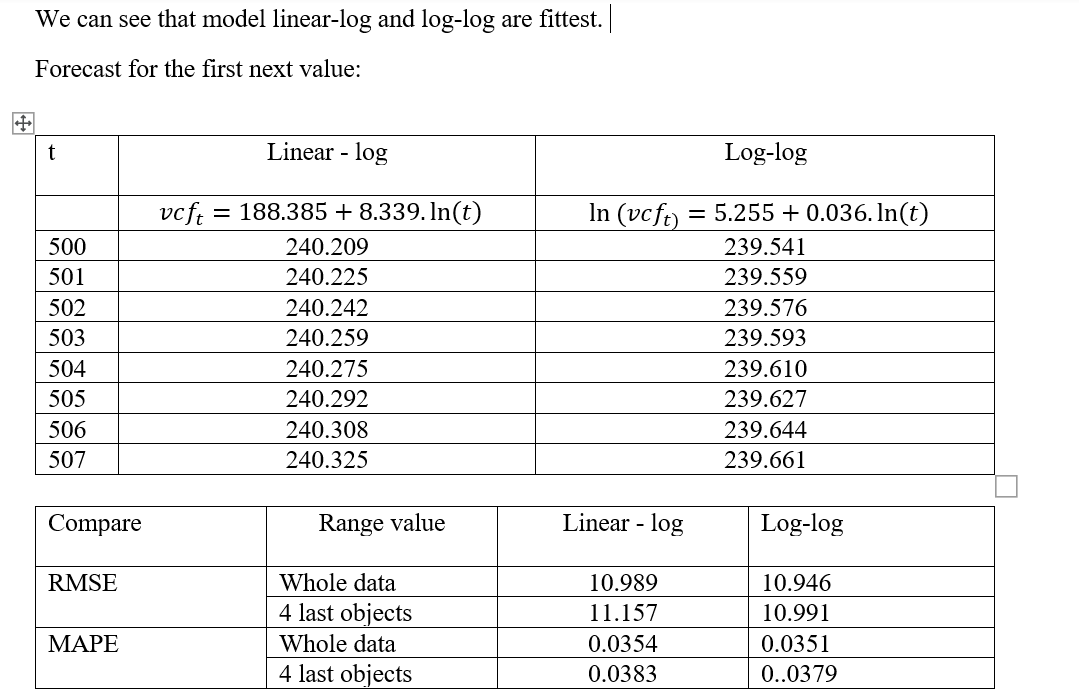
Chart, histogram

Description automatically generated

| Chart, histogram  Description automatically generated | Chart, line chart, histogram  Description automatically generated |
| --- | --- |

The distribution of price for VCF is right-skewed because it’s longer on the right side of its peak. This means the outliers of the distribution curve are further out towards the right and closer to the mean on the left





1. Stock price

a. Plot

The following time series data that represents the stock price for VCF during 2 years’ periods

Graphical user interface, application

Description automatically generatedClearly the prices are trending upwards and Nonstationary over time, but there also appears to be a cyclical or seasonal trend in the data, which can be seen by the tiny “hills” that occur over time.  
To gain a better view of this cyclical trend, we can detrend the data. In this case, this would involve removing the overall upward trend over time so that the resulting data represents just the cyclical trend. One way to detrend time series data is to simply create a new dataset where each observation is the difference between itself and the previous observation.

A picture containing antenna, cabinet

Description automatically generated

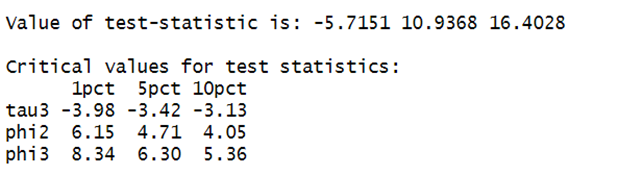
VCF is “different stationary”: I (1). It's much easier to see the seasonal trend in the time series data in this plot because the overall upward trend has been removed.

b. Unit root test for price of VCF

Test with trend and constant

Text

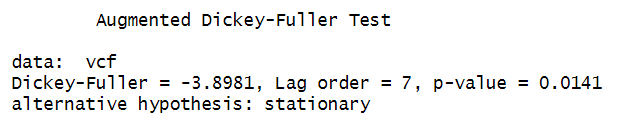
Description automatically generated



Test for significant of trend:

🡪 Reject Ho 🡪

Test for unit root



P\_val = 0.014 < 0.05 but > 0.01 🡪 Not Reject Ho 🡪

c. Unit root test for difference series

Text

Description automatically generated

Text

Description automatically generated

🡪 Reject Ho.

Mean that time series does not have a unit root, meaning it is trend stationary.

We can conclude that :

4. ACF & PACF to determine for ARMA of different series

Autocorrelation function (ACF). At lag *k*, this is the correlation between series values that are *k* intervals apart. A picture containing application

Description automatically generated

Partial autocorrelation function (PACF). At lag *k*, this is the correlation between series values that are *k* intervals apart, accounting for the values of the intervals between.

Chart

Description automatically generated

The *x* axis of the ACF plot indicates the lag at which the autocorrelation is computed; the *y* axis indicates the value of the correlation (between −1 and 1). For example, a spike at lag 1 in an ACF plot indicates a strong correlation between each series value and the preceding value, a spike at lag 2 indicates a strong correlation between each value and the value occurring two points previously, and so on.

We can observe that with ACF, the histogram has 6nd order lag and 3nd order lag PACF. The ADF test shows that the series of first difference logarithms of stock price is stationary, so we define the model ARIMA (p, d, q) suitable for prediction as ARIMA (6,1,3).

Try Criteria AIC

AIC (6) = 0.000000 🡪 min 🡪 Model: AR(6)A picture containing text

Description automatically generated

3.5. Estimate ARIMA (6,1,3) model

Text

Description automatically generated

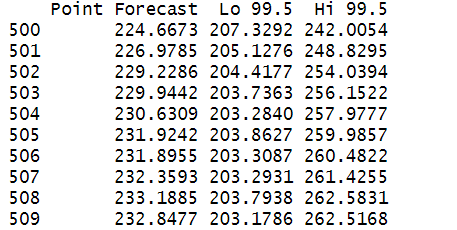
AIC=3241.26; RMSE = 6.098156; MAPE = 1.667561

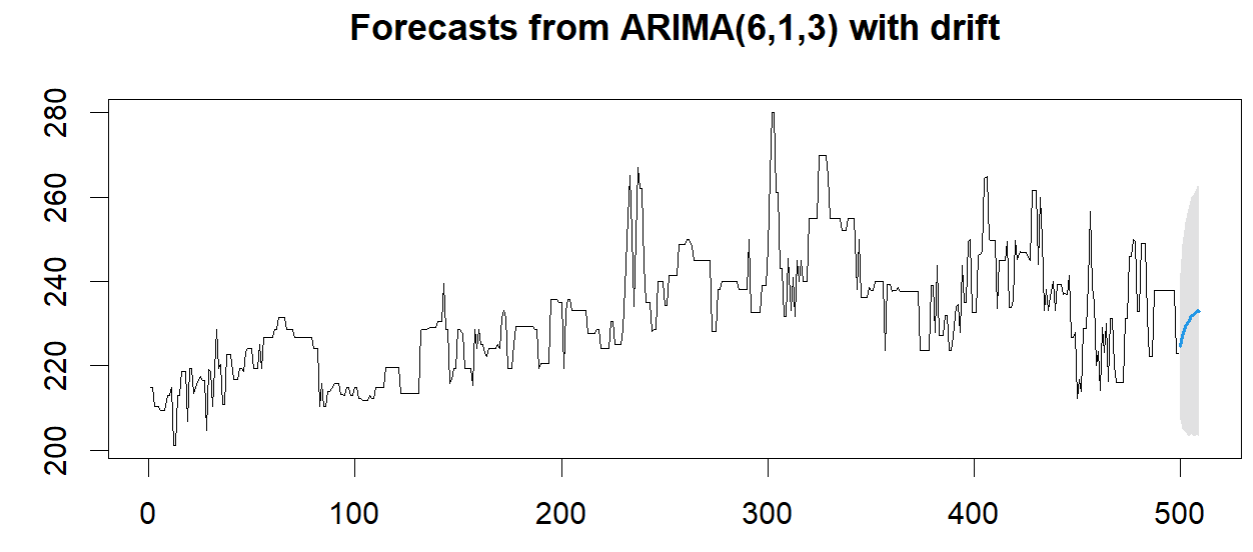
Compare ARIMA (6,1,2) model if smaller AIC better model

AIC=3244.71; RMSE = 6.152825; MAPE = 1.680848

From the result, it is clear that model ARIMA (6,1,3) is better than model ARIMA (6,1,2).

3.6. Forecast for the first 10 observations in 2023



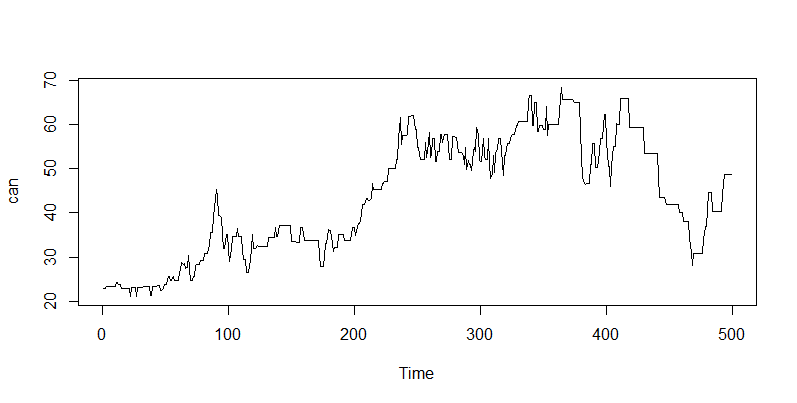


II. Hoàng Diệu Linh

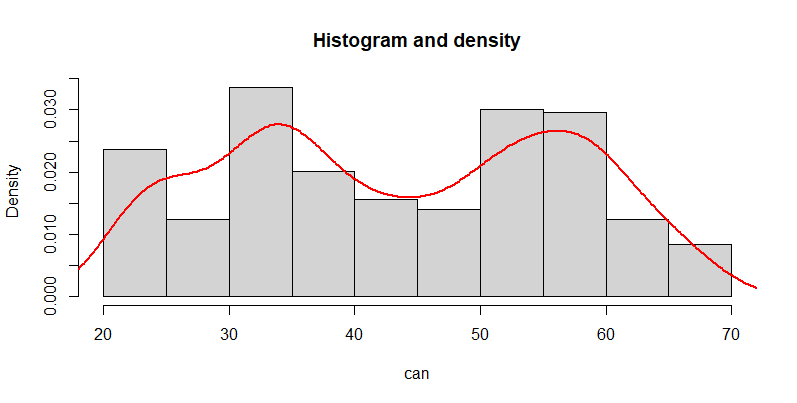
1. CTCP Đồ hộp Hạ Long (HNX: CAN)

Established in 1957, formerly known as Ha Long Canned Fish Factory, Ha Long Canned Joint Stock Company (Halong Canfoco) is considered as one of the first real canned food manufacturers in Vietnam. Today, along with the strong development of the country, Halong Canfoco is one of the first companies listed on the stock market with nearly 1,000 employees, 2 factories. The company's products are diverse from canned products such as: fish, meat, vegetables and fruits, pasteurized sausages to frozen products such as frozen sausages, spring rolls or seafood spring rolls. . The company's products are present in all provinces and cities and are exported to nearly every continent, from Europe, Asia, to the Middle East, Africa…

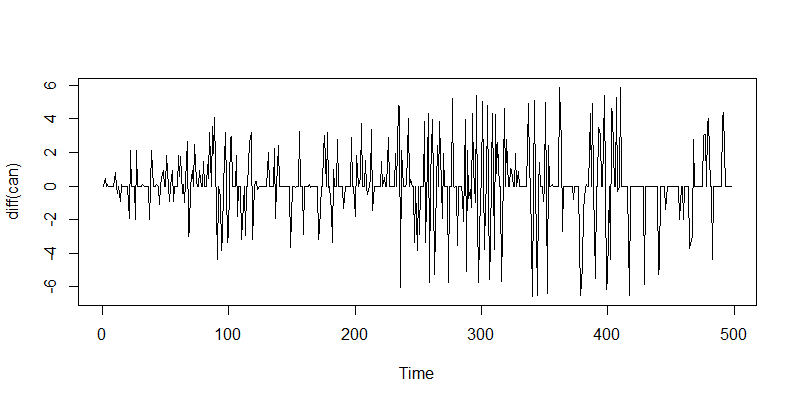
1. Finance series



The series shows a very distinct upswing from the first 400 observations; the remaining data show a very clear downturn. In general, the series tends to be volatile and unstable. Moreover, the storyline does not adequately depict the seasonal component.

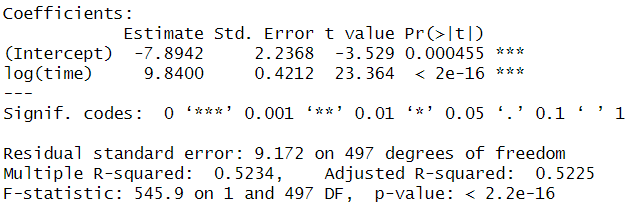


After plotting the histogram of the series, we can see that the above series is a multivariate gaussian distribution



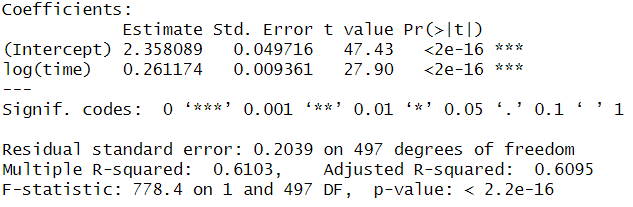
After the difference of series, we have a series that oscillates around the number 0. By subtracting the difference, we may deduce that we have a stationary series.

1. Models to forecast 8 observations in 2023
2. Linear - log



We have model:

1. Log - log



We have model:

From model we can calculate the first 8 observations:

| t | CAN ( Linear – log) | CAN (Log – log) |
| --- | --- | --- |
| 500 | 53.257 | 53.58 |
| 501 | 53.277 | 53.608 |
| 502 | 53.296 | 53.636 |
| 503 | 53.316 | 53.664 |
| 504 | 53.335 | 53.692 |
| 505 | 53.55 | 53.719 |
| 506 | 53.374 | 53.747 |
| 507 | 53.394 | 53.775 |

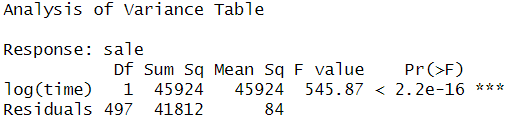
1. Compare models:

Model linear - log have: RMSE = 9.154 and MAPE = 0.196

Model log - log have: RMSE = 8.9 and MAPE = 0.175

Because both RMSE and MAPE of log - log model is smaller than linear - log model, so log - log can predict better.

Using ANOVA to test

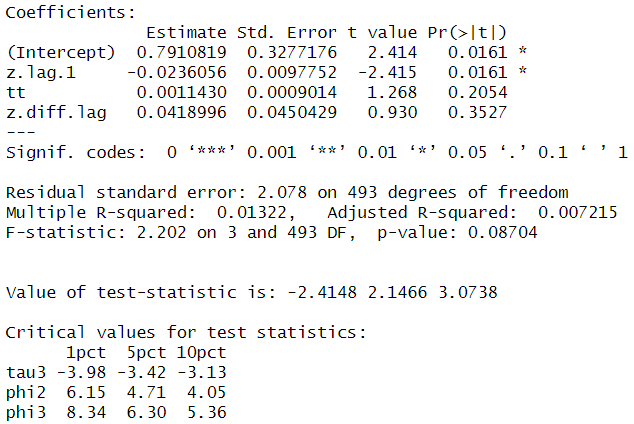


As you can see, the result shows a Df of 1 (indicating that the more complex model has one additional parameter), and a very small p-value (< .001). This means that the model log - log did lead to a significantly improved fit over the model linear - log.

1. ARIMA test for CAN
2. Unit root test

To verify that the series is stationary, we employ unit root testing. To conduct this test, we shall make use of the Dickey-Fuller test.

Test with CAN

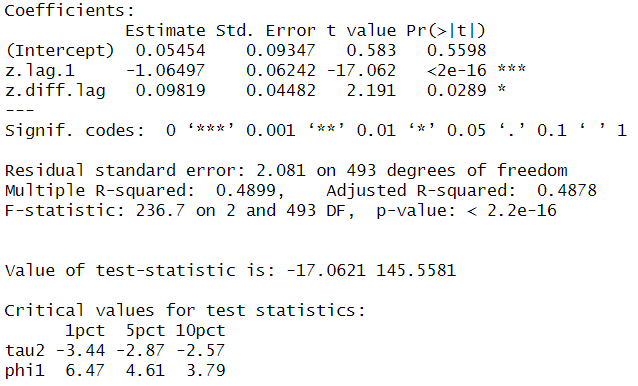


→ Not reject Ho → Can has unit root → Not stationary

After testing unit root for CAN series, we have found out that CAN has unit root, which means that the series is not stationary.

We must turn the series into a stationary series in order to apply the series to the model. We obtain a stationary series—one without trend factors—by taking the series' difference once. The values of the series are centered on the series mean.

Test with difference of CAN

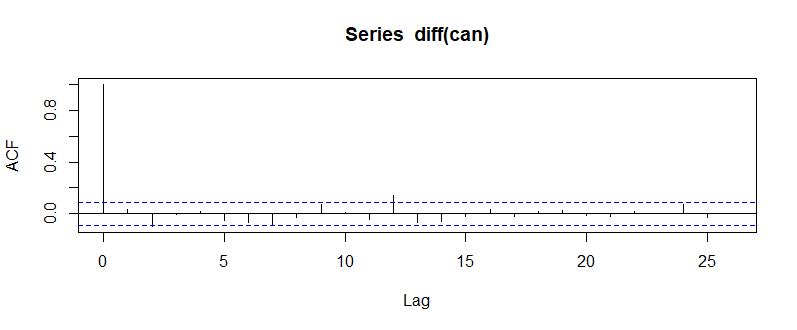


→ Reject Ho → No unit root → Stationary

After testing unit root for the difference of CAN series, We have found out that the difference of CAN does not have unit root, which means that the series is stationary. So that we will use the difference of CAN for the ARIMA model.

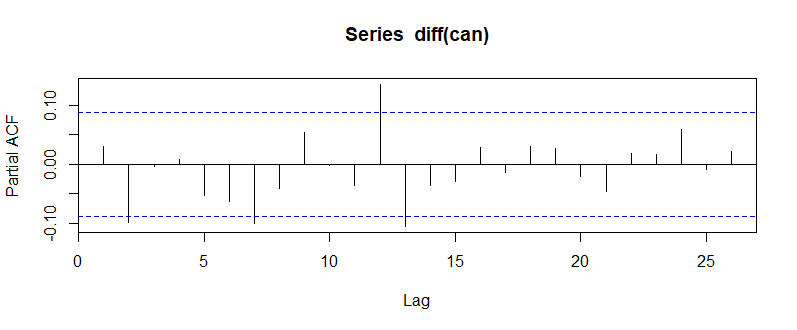
b. ACF & PACF

ACF: explains how the present value of a given time series is correlated with the past. ACF plot is a bar chart of coefficients of correlation between a time series and its lagged values. From the ACF plot we can identify the order of MA for the ARIMA model.



→ From the graph, we can conclude that MA = q = 12

PACF: is the partial autocorrelation function that explains the partial correlation between the series and lags itself. From the PACF plot we can identify the order of AR for the ARIMA model.

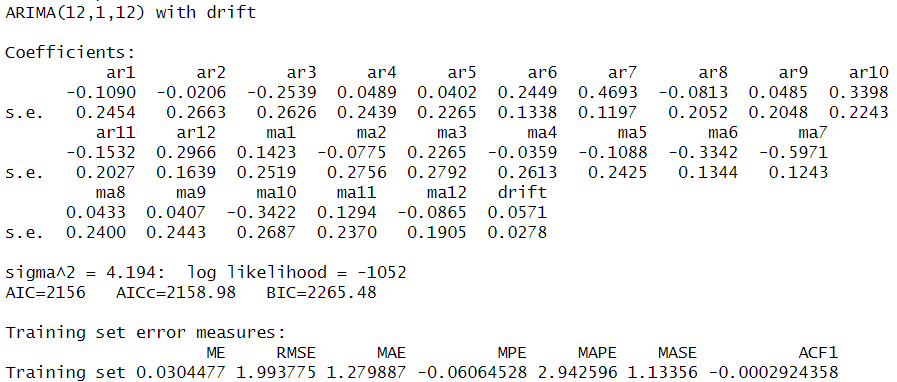


→ From the graph, we can conclude that AR = p = 12

→ Our ARIMA model is ARIMA (12,1,12)

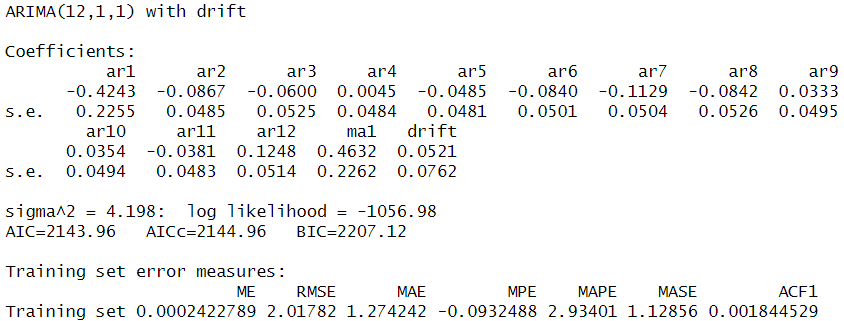
c. Estimate ARIMA model

ARIMA model (12,1,12)



→ AIC = 2156, RMSE = 1.993775, MAPE = 2.942596

ARIMA model (12,1,1)

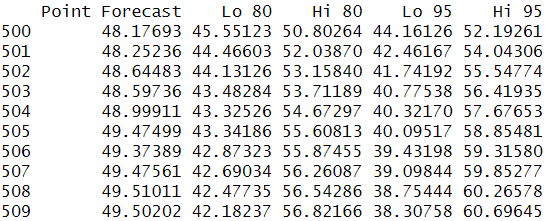


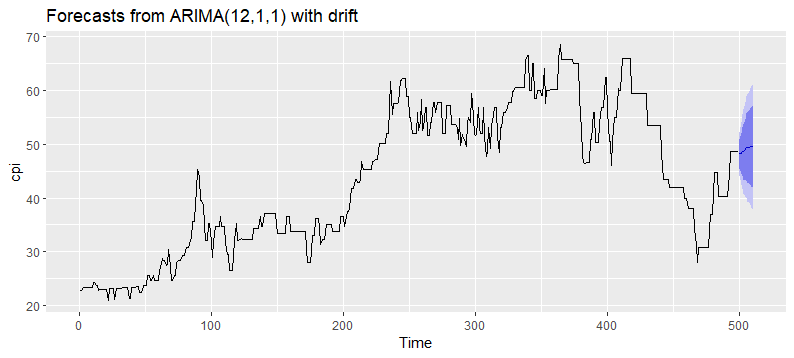
→ AIC = 2143.96, RMSE = 2.01782, MAPE = 2.93401

→ AIC of model ARIMA (12,1,12) higher than model ARIMA (12,1,1) → Model ARIMA (12,1,1) has better performance.

Thus, we will select the model ARIMA (12,1,1) to predict the subsequent 10 observations.

d. Forecast for the first 10 observations in 2023





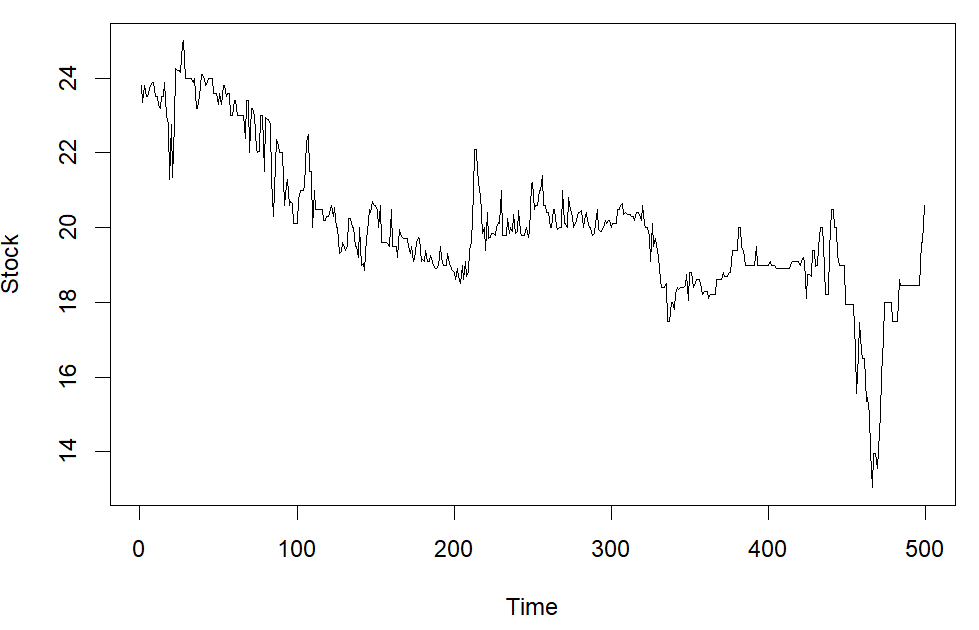
III. Nguyen Thi Thanh Hien

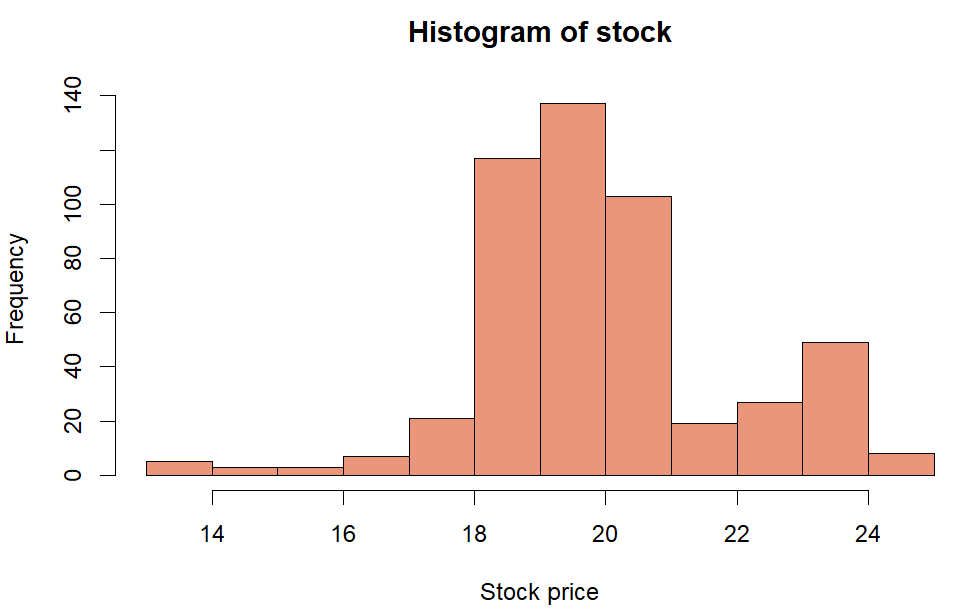
1. Chuong Duong Beverages Joint Stock Company (CDBECO)

Chuong Duong Beverages Joint Stock Company (CDBECO) is a Vietnam-based manufacturer of beverages. Its products include soft drinks, fruit drinks, mild wine and purified drinking water. The Company also offers bottling, packaging and related services to other beverage producers, as well as trades supplies and materials for the beverage industry. Moreover, it is involved in real estate trading and the offering of real estate brokerage services. As of December 31, 2012, the Company was a 51%-owned subsidiary of Saigon Beer - Alcohol - Beverage Joint Stock Corporation (SABECO). As of the same date, it had three branches located in Ho Chi Minh City, Binh Duong Province and Vinh Long Province, Vietnam.

2. Finance Series

This data shows the stock price of Chuong Duong Beverages Joint Stock Company in 2021 and 2022.





The distribution of the SCD series looks like the normal distribution except that it has a large peak at one tail. Usually this is caused by faulty construction of the histogram, with data lumped together into a group labeled “greater than.”

Forecast by using Linear - Linear model:

| 501 | 17.69 |
| --- | --- |
| 502 | 17.68 |
| 503 | 17.67 |
| 504 | 17.66 |
| 505 | 17.65 |
| 506 | 17.64 |
| 507 | 17.63 |
| 508 | 17.62 |

Forecast by using Linear - Log model

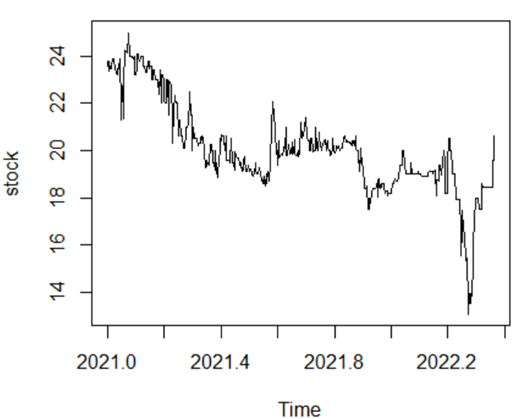
| 501 | 18.436 |
| --- | --- |
| 502 | 18.432 |
| 503 | 18.429 |
| 504 | 18.426 |
| 505 | 18.423 |
| 506 | 18.420 |
| 507 | 18.417 |
| 508 | 18.414 |

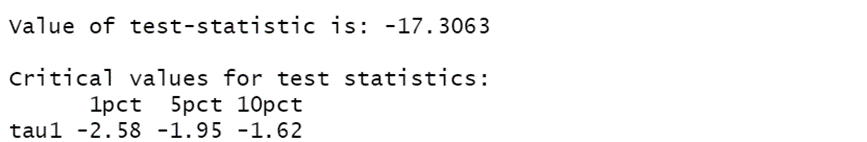
3. Stock Price

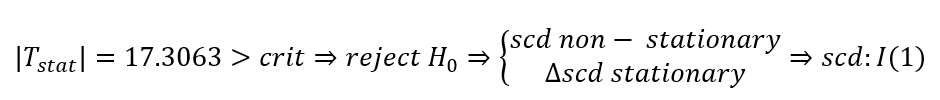
This data shows the stock price of Chuong Duong Beverages Joint Stock Company in 2021 and 2022.

a. Unit Root Test

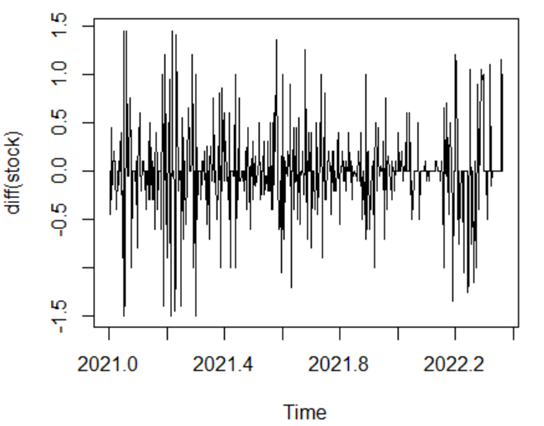
Unit root test for SCD

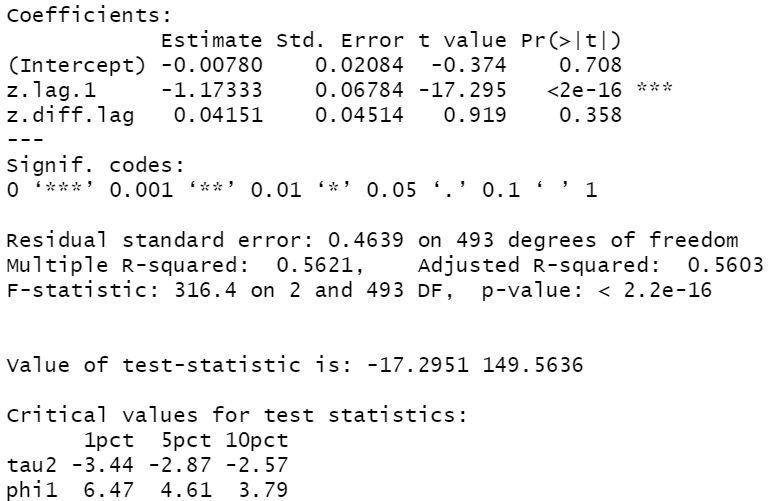


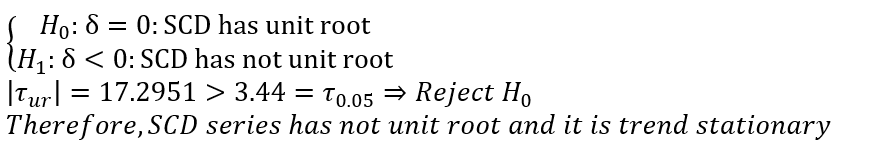




Unit Root Test for Different Series

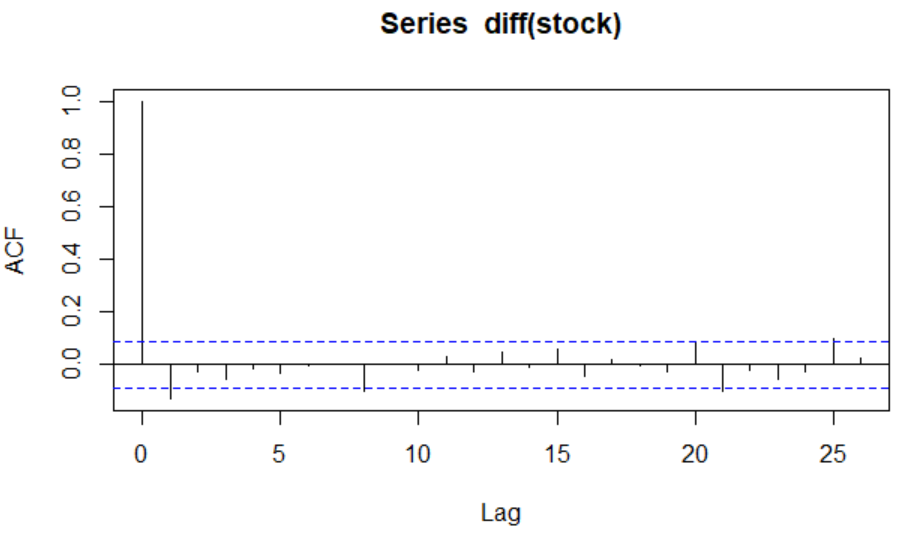






b. ACF and PACF

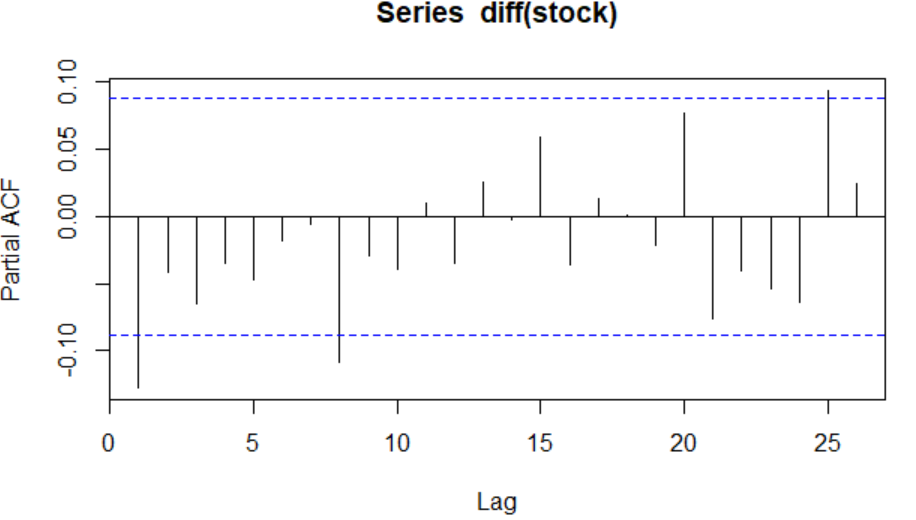
ACF: Autocorrelation is the correlation between two values in a time series. In other words, the time series data correlate with themselves—hence, the name. We talk about these correlations using the term “lags.” Analysts record time-series data by measuring a characteristic at evenly spaced intervals—such as daily, monthly, or yearly. The number of intervals between the two observations is the lag. The autocorrelation function (ACF) assesses the correlation between observations in a time series for a set of lags.



The x-axis corresponds to the different lags of the residuals. Whereas the y-axis shows the correlation of each lag. Finally, the dashed blue line represents the significance level.

After the lag-0 correlation, the subsequent correlations drop quickly to zero and stay (mostly) between the limits of the significance level (dashed blue lines). Therefore, we can conclude that the residuals of this model meet the assumption of no autocorrelation.

PACF: The partial autocorrelation function is similar to the ACF except that it displays only the correlation between two observations that the shorter lags between those observations do not explain. The partial autocorrelation function (PACF) is more useful during the specification process for an autoregressive model.

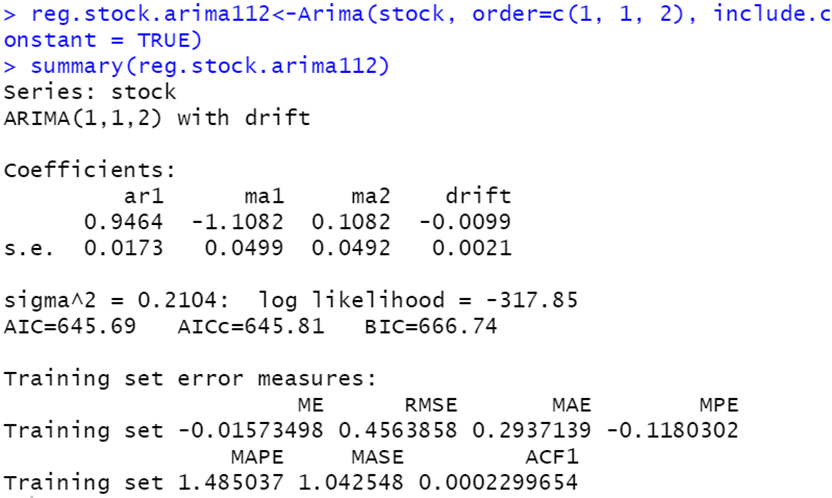


c. ARIMA (1, 1, 2) model

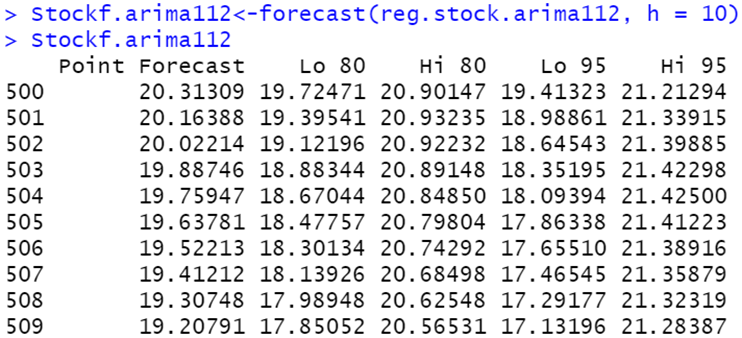
ARIMA is a method for forecasting or predicting future outcomes based on a historical time series. It is based on the statistical concept of serial correlation, where past data points influence future data points.

An ARIMA model can be understood by outlining each of its components as follows:

* Autoregression (AR): refers to a model that shows a changing variable that regresses on its own lagged, or prior, values.
* Integrated (I): represents the differencing of raw observations to allow the time series to become stationary (i.e., data values are replaced by the difference between the data values and the previous values).
* Moving average (MA): incorporates the dependency between an observation and a residual error from a moving average model applied to lagged observations.



d. Forecast for the first 10 observations in 2023



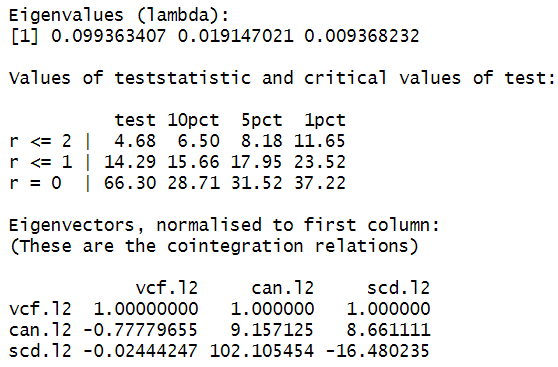
B. Group

I. Test for cointegration

Cointegration: A relationship between two non-stationary time series is possible. Two-time series that are impacted by the same underlying source, such as market forces, may “move together” over time. This relationship is referred to as co-integration. We will use Johansen test to do this test

1. Johansen test

a. Johansen test using “trace” criteria



The number of cointegrate is 0 (r = 0)

🡪 Reject Ho 🡪 Number of cointegrate > 0

The number of cointegrate is 1 (r = 0)

🡪 Not reject Ho 🡪 Number of cointegrate = 1

🡪 There are 1 cointegrate between 3 series

b. Johansen test using “eigen value” criteria

Text

Description automatically generated

The number of cointegrate is 0 (r = 0)

🡪 Reject Ho 🡪 Number of cointegrate > 0

The number of cointegrate is 1 (r = 0)

🡪 Not reject Ho 🡪 Number of cointegrate = 1

🡪 There are 1 cointegrate between 3 series

After using Johansen test with two criteria “trace” and “eigen value” we got the same results that there is one relationship among 3 series.

2. Unit root test

Based on the results of the individual part, we have the result that all 3 series have unit root, and the difference of 3 series has no unit root.

3. Test cointegrate

Series are cointegrated when their trends are not too far apart and are in some sense similar. This vague statement, though, can be made precise by conducting a cointegration test, which tests whether the residuals from regressing one series on the other one are stationary. If they are, the series are cointegrated. Thus, a cointegration test is in fact a Dickey-Fuller stationarity test on residuals, and its null hypothesis is of non-cointegration.

Text

Description automatically generated

Stationary residual test

Text

Description automatically generated

Text

Description automatically generated

In absolute terms the test statistic value of 7.043 is higher than of 3 critical values (2.58, 1.95 and 1.62). Because residues are stationary, then series are coinciding, and it means that there exists some long term relationship between variables.

4. Error correction model

Cointegration implies that time series will be connected through an error correction model. The error correction model is important in time series analysis because it allows us to better understand long-run dynamics. Additionally, failing to properly model cointegrated variables can result in biased estimates. The error correction model:

- Reflects the long-run equilibrium relationships of variables.

- Includes a short-run dynamic adjustment mechanism that describes how variables adjust when they are out of equilibrium.

- Uses adjustment coefficients to measure the forces that push the relationship towards long-run equilibrium.

Text

Description automatically generated

Correction coefficient is -0.2156. So we can conclude that this is a weak correction. The adjusted speed of the VCF series is 21.6%. It means that after each period, VCF adjusts 21.6% of error.

II. VECTOR AUTOREGRESSIVE (VAR)

1. Granger causality test

The Granger causality test is a statistical hypothesis test for determining whether one time series is a factor and offer useful information in forecasting another time series.

Let  and  be stationary time series. To test the null hypothesis that  does not Granger-cause  , one first finds the proper lagged values of   to include in an univariate autoregression of :

Next, the autoregression is augmented by including lagged values  :

One retains in this regression all lagged values of that are individually significant according to their t-statistics, provided that collectively they add explanatory power to the regression according to an F-test (whose null hypothesis is no explanatory power jointly added by the *'*s). In the notation of the above augmented regression, *p* is the shortest, and *q* is the longest, lag length for which the lagged value of is significant.

The null hypothesis that  does not Granger-cause is accepted if and only if no lagged values of are retained in the regression.

| Text, letter  Description automatically generated | Text, letter  Description automatically generated |
| --- | --- |

The hypothesis:

p-value = 0.5591, Null hypothesis is not rejected hence

And

explains to and vice versa:

The hypothesis:

p-value = 0.4055, Null hypothesis is not rejected hence

And

2. VAR (1) Model

VAR models (vector autoregressive models) are used for multivariate time series. The structure is that each variable is a linear function of past lags of itself and past lags of the other variables.

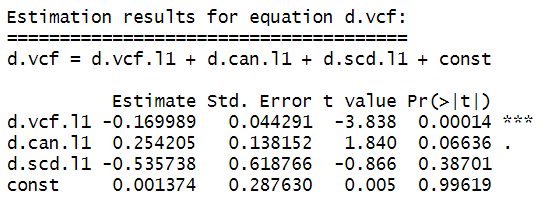
The vector autoregressive model of order 1, denoted as VAR(1), each variable is a linear function of the lag 1 values for all variables in the set.

Diagram

Description automatically generated with low confidence

a. Estimate

var1 <- VAR (data1, p =1, type = "const")



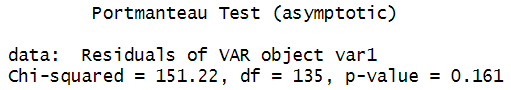
Text

Description automatically generated

Text

Description automatically generated

b. Serial correlation of Residual test



P\_value = 0.161, Null hypothesis is not rejected hence

c. Forecast with VAR (1)

Table

Description automatically generatedText

Description automatically generated

Graphical user interface, text, application, Word

Description automatically generated

4. Impulse response function

Impulse response coefficients

Text

Description automatically generated

Text

Description automatically generatedText

Description automatically generated

Impulse responses are best represented in graphs showing the responses of a VAR endogenous variable in time.

Graphical user interface, application

Description automatically generated

The interpretation here is straightforward: an impulse (shock) to DC at time zero has large effects the next period, but the effects become smaller and smaller as the time passes. The dotted lines show the 95 percent interval estimates of these effects. The VAR function prints the values corresponding to the impulse response graphs.

5. Forecast error variance decomposition – fevd

The forecast error variance decomposition is based upon the orthogonalized impulse response coefficient matrices and allow the user to analyze the contribution of variable j to the h-step forecast error variance of variable k. Forecast error variance decomposition (FEVD) is an econometric tool used by many economists in the vector autoregression (VAR) context for assessing the driving forces of business cycles.

Table

Description automatically generatedText

Description automatically generated with medium confidence

Graphical user interface, application

Description automatically generated

**CONCLUSION**

Time series are constructed using data measured over time at evenly spaced intervals. This paper mainly introduces the basic concepts of time series and time series processing models. The main method for forecasting that time is the ARIMA model. However, the ARIMA model is only suitable for stationary and linear time series, so time series with fast variation or short historical data series give inaccurate results. The time series in economics, due to the characteristics of economic development, depends a lot on different factors, so it has many variations and is nonlinear. Therefore, the ARIMA model cannot handle well in the economic field. In addition, in our report, we also use both cointegration and VAR models to test whether the series are correlated with each other and to predict the factors in the past that will affect the present and the future.